D108–D110 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj143

TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes

V. Matys*, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender

BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany

Received September 15, 2005; Revised and Accepted October 27, 2005

ABSTRACT

The TRANSFAC® database on transcription factors, their binding sites, nucleotide distribution matrices and regulated genes as well as the complementing database TRANSCompel® on composite elements have been further enhanced on various levels. A new web interface with different search options and integrated versions of Match™ and Patch™ provides increased functionality for TRANSFAC®. The list of databases which are linked to the common GENE table of TRANSFAC® and TRANSCompel® has been extended by: Ensembl, UniGene, EntrezGene, HumanPSD™ and TRANSPRO™. Standard gene names from HGNC, MGI and RGD, are included for human, mouse and rat genes, respectively. With the help of InterProScan, Pfam, SMART and PROSITE domains are assigned automatically to the protein sequences of the transcription factors. TRANSCompel® contains now, in addition to the COMPEL table, a separate table for detailed information on the experimental EVIDENCE on which the composite elements are based. Finally, for TRANSFAC®, in respect of data growth, in particular the gain of *Drosophila* transcription factor binding sites (by courtesy of the *Drosophila* DNase I footprint database) and of *Arabidopsis* factors (by courtesy of DATF, Database of *Arabidopsis* Transcription Factors) has to be stressed. The here described public releases, TRANSFAC® 7.0 and TRANSCompel® 7.0, are accessible under http://www.gene-regulation.com/pub/databases.html.

INTRODUCTION

For a better understanding of almost all life processes, a deeper knowledge of gene regulation seems indispensable. TRANSFAC® (1–3) and TRANSCompel® (3,4) are among those databases which have been contributing for years to sight and order the published data on eukaryotic gene transcription regulation and, by doing so, to make the available data applicable for analysis and predictions. The primary data in the two databases (e.g. DNA-binding sites in TRANSFAC®, composite elements in TRANSCompel®) are based on experimental evidence. These data are extracted by curators from peer-reviewed papers. The curators search the scientific literature for suitable data, which are then entered via an input client, making use of controlled vocabulary and various automated functions, into a relational database, from which flatfile releases are generated from time to time. Collection of these data in a structured form allows us to deduce—via comparison and classification—secondary or so-called meta-data (e.g. nucleotide distribution matrices, factor classification). Both types of data, the primary as well as the secondary data, can then serve for (sequence-based) predictions by certain programs, e.g. Match™ (5) (for matrix-based transcription factor binding site searches), Patch™ (for pattern-based transcription factor binding site searches) and P-Match™ (6) (for a mixture of matrix- and pattern-based binding site searches). (These programs are all available on the same server as the herein described databases.)

Content of TRANSFAC® and TRANSCompel®

The primary data of TRANSFAC® are stored in the three tables FACTOR, SITE and GENE for information on transcription factors, their binding sites and regulated genes, respectively. Besides genomic binding sites, the SITE table contains also so-called artificial binding sites, which are mostly sites from random oligonucleotide selection assays, and IUPAC consensus sequences. Nucleotide distribution matrices, which are derived from a collection of binding sites for a particular factor are stored in the MATRIX table, while the CLASS table groups the transcription factors according to their DNA-binding domains. In addition to this CLASS table, the factor entries are linked to the respective nodes in a classification hierarchy (7). In a sixth table (CELL), cell lines

*To whom correspondence should be addressed. Tel: +49 5331 8584 28; Fax: +49 5331 8584 70; Email: vma@biobase.de

Table 1. Number of entries in the tables of TRANSFAC® 7.0 and TRANSCompel® 7.0

Table	TRANSFAC® Rel. 7.0
FACTOR	6133
Homo sapiens	1040
Mus musculus	765
D.melanogaster	233
A.thaliana	1751
S.cerevisiae	368
SITE	7915
MATRIX	398
GENE (all entries)	2397
H.sapiens	608
M.musculus	417
D.melanogaster	145
A.thaliana	115
S.cerevisiae	195
GENE (entries with SITE links)	1504
CLASS	50
CELL	1307
	TRANSCompel® Rel. 7.0
COMPEL (composite elements)	322

and other kinds of factor sources, which were used for detection of a binding site/binding activity, are stored (Table 1).

While TRANSFAC® deals essentially with single factor–site interactions, the focus of TRANSCompel® is on so-called composite elements, consisting of two (or more) neighboring binding sites, characterized by synergistic or antagonistic effects between the two transcription factors binding to them. They are, thus, the smallest units of combinatorial transcriptional regulation (Table 1).

Recent changes in the database structure

In the TRANSFAC® GENE table a new field has been introduced for the inclusion of information on the regulation of gene expression, especially when this information cannot be assigned (yet) to a particular binding site. The CELL table, which contains entries of cell lines or other factor sources, lists now all SITE entries, for which binding activity was shown under the given conditions. The factor CLASS entries have been linked to the respective nodes in the hierarchical factor classification. The links from GENE entries (TRANSFAC®) to composite elements (in TRANSCompel®) are no longer listed among other database links, but are given now subsequent to the listed binding sites, and, as for those, the positions of the composite elements within the gene (usually relative to the transcription start site, TSS) are given. The structure of TRANScompel® has been fundamentally changed. The database consists now of two tables. The COMPEL table contains general information about the composite elements including sequence, positions, gene, names of cooperating transcription factors as well as a brief list of the experimental evidence, while detailed information about the experimental evidence, confirming physical and functional interactions between the corresponding transcription factors, can be found in the EVIDENCE table.

Linking to other databases

Linking to other databases has been extended. The FACTOR and GENE entries of TRANSFAC® have been linked to the respective protein reports in HumanPSD™ (8). GENE entries from human, mouse and rat have been linked to the promoter database TRANSPRO™ [see the paper about TiProD (9)], and the corresponding SITE entries were also mapped to the promoter sequences in TRANSPRO™, where absolute genomic positions are given for the promoter sequences. These sequences in TRANSPRO™ reach from 10 000 nt upstream to 1000 nt downstream relative to the accepted 'virtual TSSs', which are derived by weighting documented TSSs from different resources (9). FACTOR–SITE–GENE links are mirrored as '*trans*-regulations' (FACTOR→GENE) in TRANSPATH® (10,11), where these are incorporated into the overall regulation network of the cell. For yeast (*Saccharomyces cerevisiae*) genes, now standard open reading frame names are given under synonyms. For human, mouse and rat genes (and factors) HGNC (12), MGI (13) and RGD (14) gene symbols are given, respectively. (The HGNC, MGI and RGD gene symbols appear in the GENE table under external database links and in the FACTOR table alongside the GENE link.) Further, new links to Ensembl (15) and UniGene (16), as well as Affymetrix probe set IDs were added to the GENE table, and the links to LocusLink were changed into EntrezGene (17) links and expanded from human to mouse and rat, as well as—to a smaller extend—other organisms.

Automatic factor domain assignment

For each release protein sequences are analyzed with InterProScan (18). From the databases integrated by InterPro, Pfam (19) and SMART (20), as well as PROSITE (21) models corresponding to low-complexity regions are selected. The automatically assigned domains, which are linked to the corresponding Pfam, SMART and PROSITE entries, are meant to complement the manually annotated domains, many of which are based on functional studies reported in the original literature.

From *Arabidopsis* to *Drosophila*

Besides a general data increase, with major focus on human, mouse, rat and other vertebrate organisms, especially the amount of *Arabidopsis thaliana* and *Drosophila melanogaster* data has been increased for TRANSFAC® 7.0. This was accomplished in particular by import of 1440 factor entries from DATF, Database of *Arabidopsis* transcription factors [http://datf.cbi.pku.edu.cn/ (22)], and 899 genomic site entries from the *Drosophila* DNase I footprint database [http://www.flyreg.org/ (23)]. The imported data are referenced accordingly and are linked to the respective databases, from which they were derived. In addition, the *Drosophila* gene entries linked to the newly imported sites contain pointers to FlyBase (24) and EntrezGene (17). Identifiers used by Ensembl which are synonyms in FlyBase and EntrezGene (e.g. CG3481) were introduced as synonyms of the gene name and were used for mapping during the import procedure.

New web interface

TRANSFAC® and the programs Match™ and Patch™, for transcription factor binding site searches, are now combined under a common web interface. In addition to the 'one table search' the new search engine has a search mode for

simultaneous search in all tables. The 'one table search' contains now the possibility to combine searches in up to three fields at the same time or to make batch search and further options. The user can also choose the output fields, which are to be displayed in the result list. Each search can be stored and refined later on or the stored search result (list of accession numbers) can serve as input for a batch search in another table. Finally, on the basis of the result from a search in the SITE or MATRIX table, 'profiles' (sets of sites or matrices) can be created, which can be used by the integrated version of Patch™ or Match™, respectively, for sequence analyses.

AVAILABILITY

The described TRANSFAC® 7.0 and TRANSCompel® 7.0 releases as well as the programs Match™, Patch™ and P-Match™ are all freely available for online use by users from non-profit organizations at http://www.gene-regulation.com/pub/databases.html#transfac, http://www.gene-regulation.com/pub/databases.html#transcompel and http://www.gene-regulation.com/pub/programs.html, respectively.

ACKNOWLEDGEMENTS

We like to thank Prof Jingchu Luo, Dr Anyuan Guo and colleagues from Peking University, Center for Bioinformatics, for providing the DATF data and Dr Casey Bergman and colleagues from FlyReg, Universiy of Manchester, U.K. for the *Drosophila* footprint data, as well as Prof. Dr. Reinhard Hehl and Claudia Galuschka from the Technical University Braunschweig, Germany for their collaboration on plant data curation. Further, we like to thank all people who have been contributing over the years to the development and curation of the described databases and connected tools. Parts of the work were funded by grants of the German Ministry of Education and Research (BMBF) 'Intergenomics' (031U210B), collectively by BioRegioN GmbH and BMBF 'BioProfil' (0313092), by the European Commission under FP6-'Life sciences, genomics and biotechnology for health', contract LSHG-CT-2004-503568 'COMBIO', and by the European Commission under 'Marie Curie research training networks', contract MRTN-CT-2004-512285 'TRANSISTOR'. Funding to pay the Open Access publication charges for this article was provided by BIOBASE GmbH.

Conflict of interest statement. None declared.

REFERENCES

1. Wingender,E. (1988) Compilation of transcription regulating proteins. *Nucleic Acids Res.*, **16**, 1879–1902.
2. Matys,V., Fricke,E., Geffers,R., Gößling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
3. Kel-Margoulis,O., Matys,V., Choi,C., Reuter,I., Krull,M., Potapov,A.P., Voss,N., Liebich,I., Kel,A. and Wingender,E. (2005) Databases on gene regulation. In Bajic,V.B. and Tan,T.W. (eds), *Information Processing and Living Systems*. World Scientific Publishing Co., Singapore, Vol. 2, pp. 709–727.
4. Kel-Margoulis,O., Kel,A.E., Reuter,I., Deineko,I.V. and Wingender,E. (2002) TRANSCompel—a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.
5. Kel,A.E., Gößling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
6. Chekmenev,D.S., Haid,C. and Kel,A.E. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.*, **33**, W432–W437.
7. Wingender,E. (1997) Classification scheme of eukaryotic transcription factors. *Mol. Biol.*, **31**, 483–497.
8. Hodges,P.E., Carrico,P.M., Hogan,J.D., O'Neill,K.E., Owen,J.J., Mangan,M., Davis B.P., Brooks,J.E. and Garrels,J.I. (2002) Annotating the human proteome: the Human Proteome Survey Database (HumanPSD™) and an in-depth target database for G protein-coupled receptors (GPCR-PD™) from Incyte. Genomics. *Nucleic Acids Res.*, **30**, 137–141.
9. Chen,X., Wu,J.-m., Hornischer,K., Kel,A. and Wingender,E. (2006) TiProD: the Tissue-specific Promoter Database. *Nucleic Acids Res.*, **34**, D104–D107.
10. Schacherer,F., Choi,C., Götze,U., Krull,M., Pistor,S. and Wingender,E. (2001) The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, **17**, 1053–1057.
11. Krull,M., Voss,N., Choi,C., Pistor,S., Potapov,A. and Wingender,E. (2003) TRANSPATH®: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.*, **31**, 97–100.
12. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, **32**, D255–D257.
13. Bult,C.J., Blake,J.A., Richardson,J.E., Kadin, J.A., Eppig,J.T. and the Mouse Genome Database Group (2004) The Mouse Genome Database (MGD): integrating biology with the genome. *Nuceic Acids Res.*, **32**, D476–D481.
14. de la Cruz,N., Bromberg,S., Pasko,D., Shimoyama,M., Twigger,S., Chen,J., Chen,C.F., Fan,C., Foote,C., Gopinath,G.R. *et al.* (2005) The Rat Genome Database (RGD): developments towards a phenome database. *Nucleic Acids Res.*, **33**, D485–D491.
15. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
17. Maglott,D., Ostell,J., Pruitt,K.D., Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
18. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
19. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
20. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P., Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
21. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P., Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
22. Guo,A., He,K., Liu,D., Bai,S., Gu,X., Wei,L., Luo,J. (2005) DATF: a Database of *Arabidopsis* Transcription Factors. *Bioinformatics*, **21**, 2568–2569.
23. Bergman,C.M., Carlson,J.W., Celniker,S.E. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.
24. Drysdale,R.A., Crosby,M.A. and FlyBase Consortium (2005) FlyBase: genes and gene models. *Nucleic Acids Res.*, **33**, D390–D395.